Claims

1. A computer-implemented method (1000) for extracting dynamic content data (221) in a machine-readable format from a website (220) provided by a server wherein dynamic content data (221) relates to content data which is generated by the server in response to a request, the method comprising:

    accessing (1100) configuration data reflecting the structure of the website (220), the configuration data including at least a website specific scraping script having instructions emulating a browser to automatically perform parameterized navigation steps on the website;

    receiving (1200), via an interface (110), a data retrieval request (210) specifying the website (220) and corresponding dynamic content data (221) to be retrieved from the website (220);

    triggering (1300) execution of the scraping script to access the dynamic content data (221), wherein the scraping script has a step description including one or more dynamic parameters with data from the data retrieval request specifying the dynamic content to be retrieved;

    receiving (1400), from the website (220) in response to the scraping script execution, HTML/XML data representing the dynamic content data wherein the value of a particular dynamic parameter is the result value of a particular query;

    providing (1500) the HTML/XML data to an XPath extraction module (150), wherein the XPath extraction module is pre-configured with the website specific XPath statements in accordance with the structure of the website (220); and

    receiving (1600), from the XPath extraction module (150), machine-readable content data (222) extracted from the HTML/XML data;

    wherein the data retrieval request (210) specifying the dynamic content data (221) includes one or more dynamic parameters of the scraping script that are replaced with values from preceding server responses extracted from previous navigation steps, the replaced one or more parameters to be used for the next navigation step, and further including one or more website specific XPath statements addressing said values.

2. The method of claim 1, wherein the website requires login credentials and wherein the data retrieval request further specifies user authentication data, the method further comprising:

providing (1210) the authentication data to an account database (230); and

in case of successful authentication of the user, receiving (1220) from the account database (230) the login credentials for the authenticated user in response to a credential request.

3. The method of claim 1 or 2, wherein the data retrieval request is buffered in a request queue (115) amongst a plurality of further data retrieval requests.

4. The method of any of the previous claims, further comprising:

in case of a login requirement for the website, determining (1240) a session ID for the data retrieval request to access the dynamic content data on the website.

5. The method of any of claim 4, wherein determining (1240) a session ID further comprises:

if an open session is available for the received data retrieval request, providing (1242) the session ID of the open session;

if no open session is available for the received data retrieval request, triggering (1243) execution of a pre-configured login script in accordance with the data retrieval request, and receiving (1244), in response to the executed login script, one or more cookies as the session ID.

6. The method of any of the previous claims, wherein the one or more parameterized navigation steps are enabled by the scraping script using results from preceding server responses allowing for dynamic navigation through the website and enabling the scraping script to work without relying on predefined URLs.

7. The method of any of the previous claims, wherein the data retrieval request includes at least one dynamic parameter leading to different content (221) generated by the website (220) in response to the execution of a particular scraping script step, wherein the dynamic parameter value is not known beforehand.

8. A computer program product that when loaded into a memory of a computing device and executed by at least one processor of the computing device executes the steps of the computer implemented method according to any of the previous claims.

9. A computer system (100) for extracting dynamic content data (221) from a website (220) in a machine-readable format, the system comprising:

an interface (110) configured to access configuration data (250) reflecting the structure of the website (220), the configuration data including at least a website specific scraping script having instructions emulating a browser to automatically perform parameterized navigation steps on the website, and further to receive a data retrieval request (210) specifying the website (220) and corresponding dynamic content data (221) to be retrieved from the website (220;

a scraper module (120) configured to provide the scraping script (2050) for execution wherein the scraping script is configured to perform one or more parameterized navigation steps on the website (220) to access the dynamic content data (221);

a script module (140) configured to trigger execution of the scraping script wherein the scraping script has a step description including one or more dynamic parameters with data from the data retrieval request specifying the dynamic content to be retrieved, and to receive, from the website (220) in response to the scraping script execution, HTML/XML data representing the dynamic content data wherein the value of a particular dynamic parameter is the result value of a particular query; and

an XPath extraction module (150), wherein the XPath extraction module is pre-configured with the website specific XPath statements in accordance with the structure of the website (220) to extract machine-readable content data (222) from the HTML/XML data;

wherein the data retrieval request (210) specifying the dynamic content data (221) includes one or more dynamic parameters of the scraping script that are replaced with values from preceding server responses extracted from previous navigation steps, the replaced one or more parameters to be used for the next navigation step, and further including one or more website specific XPath statements addressing said values.

10. The system (100) of claim 9, further comprising:

a request queue (115) configured to buffer the data retrieval request (210) amongst a plurality of further data retrieval requests.

11. The system (100) of claim 9 or 10, wherein the website requires login credentials and wherein the data retrieval request (2010) further specifies user authentication data (2020), further comprising:

a session management module (130) configured to determine a session ID for the data retrieval request to access the dynamic content data on the website;

12. The system (100) of claim 11, the interface (110) being communicatively coupled with an account database (230), and being further configured to provide the authentication data (2020) to the account database (230), and in case of successful authentication of the user, to receive from the account database (230) the login credentials (2023) for the authenticated user in response to a credential request (2022).

13. The system (100) of claims 11 or 12, wherein the session management module (130) is further configured to:

check if an open session is available for the received data retrieval request;

if an open session is available, provide the session ID of the open session to the scraper module (120);

if no open session is available for the received data retrieval request, trigger the execution of a pre-configured login script (2040) by the script module (140) in accordance with the data retrieval request (2010), and provide one or more cookies (2042), received in response to the executed login script, as the session ID to the scraper module (120).

14. The system (100) of any of claims 9 to 13, wherein the computer system (100) is further configured to deploy configuration data to the computer system modules.

15. The system (100) of any of claims 9 to 14, wherein the one or more parameterized navigation steps are enabled by the scraping script using results from preceding server responses allowing for dynamic navigation through the website and enabling the scraping script to work without relying on predefined URLs.